Experts Recommendation

Jadie Zuo

April 11, 2016

Data Preparetion

We selecte the most deviated users: 100 most deviated reviewers, and 100 experts.

```
con <- load("connoi.RData")
ext <- load("extreme.RData")
mydata <- readRDS("users_50_products_100.RDS")
exp <- connoi[1:100,]
ext <- extreme[1:100,]
colnames(exp) <- c("userid","review_num","review_ave","help_num","help_score","dev")
colnames(ext) <- c("userid","review_num","review_ave","help_num","help_score","dev")
```

Prepare the node dataset: a 200 by 9 matrix with rows being the combination of the most deviated reviewrs and experts, and columns being the following features: * ID: identification number (a sequence from 1 to 200)
* userid: ID assigned by Amazon
* review_num: number of reviews created
* review_ave: average score of review
* help_num: number of helpfulness reviews by other users
* help_score: helpfulness score evaluated by other users
* dev: expertise measurement that is calculated by the deviation from his average review score to the overall review score
* type: binary variable with 1 being deviated reviewers and 2 being experts
* type.label: labels for type, extreme reviewers and experts

```
a <- load("node.RData")
node$type <- c(rep(1,100), rep(2,100))
node$type.label <- c(rep("Extreme Reviewers",100), rep("Experts",100))
node <- cbind(seq(1,200,1),node)
```

Prepare the edge dataset: a 321 by 4 matrix with rows being edges among 200 reviewers and following column factors:
* from: start point of an edge
* to: end point of an edge
* weight: number of movies that two nodes have commonly seen
* type: how strong the connection is, with 1 being the weight below 10 indicating a weak connection, 2 being the weight between 10 and 25 indicating a connection, 3 being weight above 25 indicating strong connection.

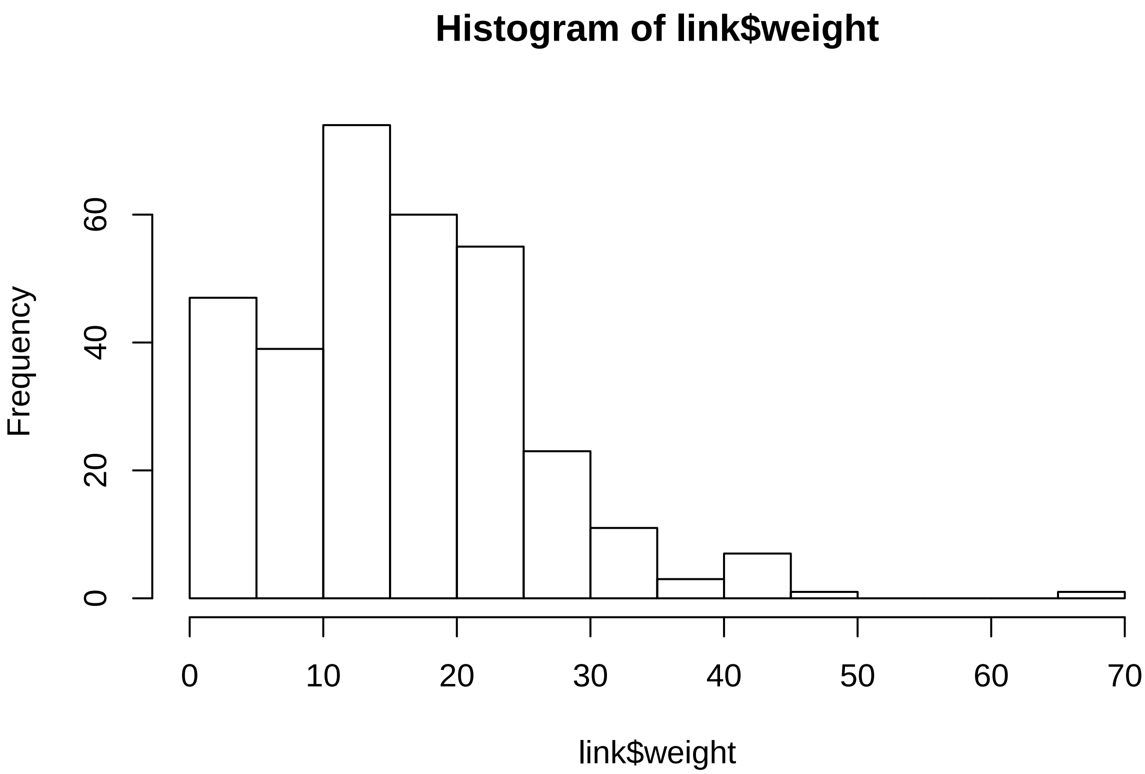
```
con <- matrix(nrow = 100,ncol = 100)
for (i in 1:100) {
  one <- mydata[which(mydata$review_userid == ext$review_userid[i]),]
  x1 = as.numeric(unique(one$product_productid))
  for (j in 1:100) {
    two <- mydata[which(mydata$review_userid == exp$review_userid[j]),]
    y1 = as.numeric(unique(two$product_productid))
    count <- length(intersect(x1, y1))
    con[i,j] <- count
```

```
  }
}

from <- c(1,1,1,1,1,2,2,2,2,3,3,3,3,4,5,6,6,6,6,6,6,6,7,7,9,9,9,9,9,9,9,9,
          9,10,10,10,10,10,10,10,11,11,11,11,11,11,11,13,13,13,13,13,13,14,
          14,14,15,16,16,16,17,17,18,18,19,19,19,20,20,21,21,21,21,22,26,26,
          26,26,26,26,27,27,28,28,28,29,29,29,29,29,29,30,32,32,32,34,34,35,
          35,35,36,36,37,37,38,38,38,38,38,38,38,38,39,39,39,39,41,41,42,42,
          42,42,42,42,42,42,43,43,43,43,43,44,44,45,45,45,45,46,46,46,46,46,
          46,46,46,46,47,47,47,49,49,49,49,49,49,50,50,50,50,50,50,50,50,50,
          50,50,50,51,51,52,52,52,54,54,54,54,54,54,54,54,55,55,56,56,56,56,
          57,57,57,57,58,58,58,59,60,60,60,61,62,63,64,64,64,64,64,66,67,67,
          68,68,69,70,70,70,70,71,71,72,72,72,72,72,72,72,72,72,73,73,73,73,
          73,74,74,76,76,76,77,77,77,78,78,78,78,79,79,80,80,80,80,81,81,81,
          81,82,82,82,82,83,83,84,84,84,84,84,84,85,85,85,86,86,87,87,87,87,
          88,89,89,89,89,89,89,90,90,91,91,91,91,92,92,92,92,93,94,94,95,96,
          96,96,96,96,96,97,98,98,98,98,98,98,98,99,99,99,99,99,99,99,100,
          100,100,100)
to <- c(34,64,82,83,95,67,75,79,83,15,35,72,94,52,92,27,30,43,59,72,6,14,57,
        11,93,22,32,54,92,94,75,76,83,67,29,67,43,64,92,83,95,83,92,70,73,86,
        87,63,90,43,92,67,75,64,99,90,93,75,50,32,75,90,93,68,61,83,91,92,64,
        17,12,92,96,90,63,40,92,61,24,54,83,70,90,66,83,92,8,14,43,59,83,100,
        29,68,90,97,83,91,83,56,43,90,75,47,83,1,32,54,67,94,90,98,92,83,92,
        75,63,21,92,55,68,82,28,75,90,22,99,16,43,64,92,100,92,67,90,75,92,42,
        93,44,58,63,39,75,89,32,92,35,92,4,38,59,60,83,20,92,95,65,99,62,10,
        67,73,56,75,92,91,90,83,53,90,83,91,53,59,43,83,88,56,75,92,48,92,56,
        75,19,63,48,49,75,83,91,90,93,85,9,26,83,65,83,99,82,43,21,75,90,91,49,
        64,83,100,83,92,67,53,75,83,90,83,8,19,42,43,64,91,92,100,68,62,6,14,
        57,47,83,90,92,32,98,82,53,91,90,83,82,83,90,41,66,32,83,53,93,90,85,
        78,62,83,55,92,90,70,12,44,75,92,45,35,72,94,83,55,68,83,24,54,53,48,
        41,75,43,19,63,56,50,39,40,43,92,92,49,75,83,79,82,92,96,73,59,69,56,
        75,92,75,64,4,63,70,73,86,87,85,90,83,34,44,92,95,11,1,42,92)
weight <- c()
for (i in 1:321){
  weight[i] <- con[from[i],to[i]]
}
```

Take a look at how "weight" is distributed:

```
hist(link$weight)
```



Creat a categorical varaible using the weight variable to describe the level of similarity between reviewers. According to the histogram, use 10 and 25 as two cut off points:

```
type <- c()
for (i in 1:321){
  if (weight[i] < 10) {
    temp <- 1
  }
  else if (weight[i] < 25) {
    temp <- 2
  }
  else {
    temp <- 3
  }
  type[i] <- temp
}
link <- data.frame(from,to,weight,type)
colnames(link) <- c("from", "to", "weight", "type")
rownames(link) <- NULL
```

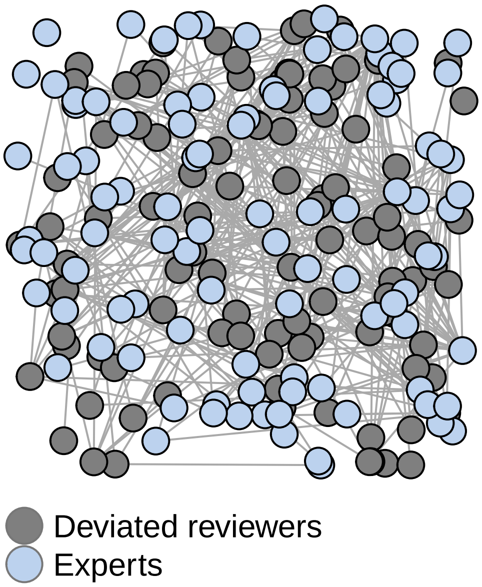
Network Plots

Network layout using igraph:

```
library(igraph)
library(RColorBrewer)
net <- graph.data.frame(link, node, directed=T)
net <- simplify(net, remove.multiple = F, remove.loops = T)
colrs <- c("gray50", "lightsteelblue2")
```

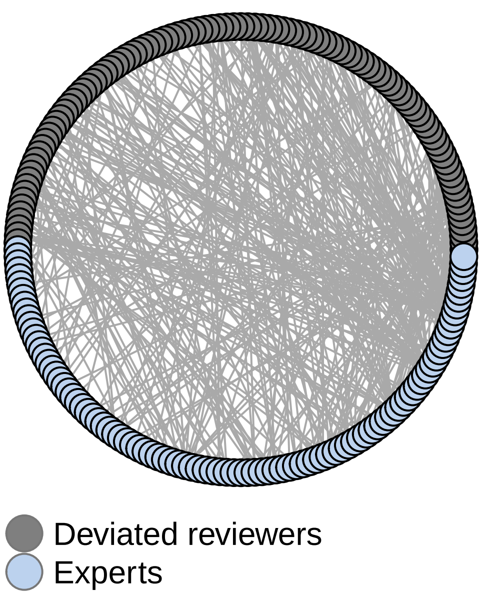
Random Network Layout

```r
plot(net, vertex.size=12, edge.arrow.size=0, edge.curved=0,vertex.color=colrs[V(net)$type],
     vertex.frame.color="black",vertex.label=NA, layout=layout.random)
legend(x=-1.1, y=-1.1, c("Deviated reviewers","Experts"), pch=21,
       col="#777777", pt.bg=colrs, pt.cex=2.5, bty="n", ncol=1)
```



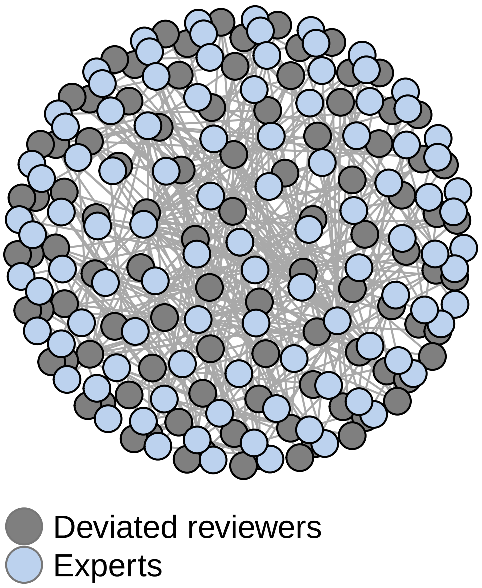
Circle Layout

```r
plot(net,vertex.size=12, edge.arrow.size=0, edge.curved=0,vertex.color=colrs[V(net)$type],
     vertex.frame.color="black",vertex.label=NA,layout=layout.circle(net))
legend(x=-1.1, y=-1.1, c("Deviated reviewers","Experts"), pch=21,
       col="#777777", pt.bg=colrs, pt.cex=2.5, bty="n", ncol=1)
```



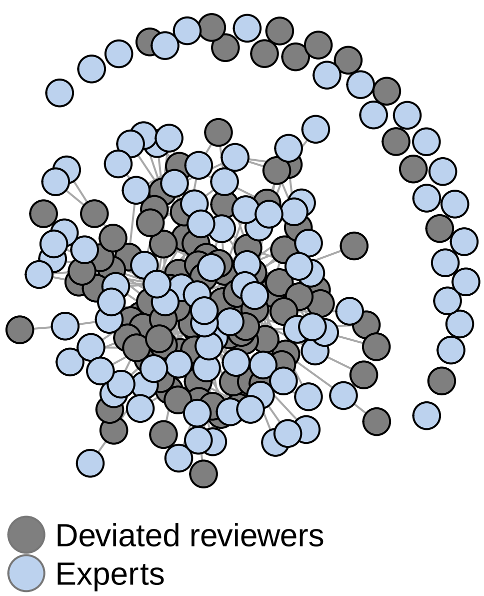
3D sphere layout:

```r
plot(net,vertex.size=12, edge.arrow.size=0, edge.curved=0,vertex.color=colrs[V(net)$type],
     vertex.frame.color="black",vertex.label=NA,layout=layout.sphere(net))
legend(x=-1.1, y=-1.1, c("Deviated reviewers","Experts"), pch=21,
       col="#777777", pt.bg=colrs, pt.cex=2.5, bty="n", ncol=1)
```



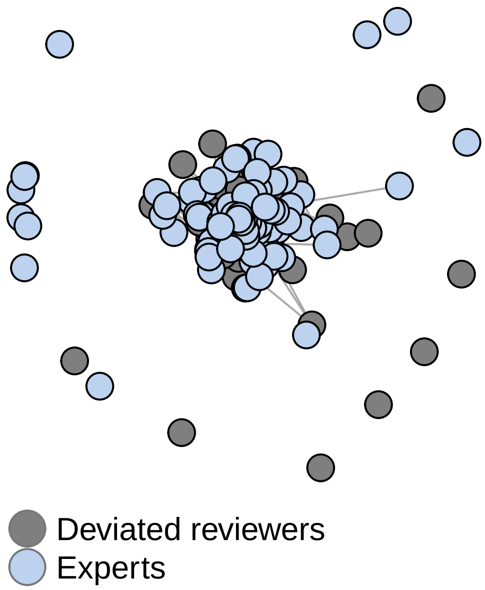
The Fruchterman-Reingold force-directed algorithm:

```r
plot(net,vertex.size=12, edge.arrow.size=0, edge.curved=0,vertex.color=colrs[V(net)$type],
     vertex.frame.color="black",vertex.label=NA,layout=layout.fruchterman.reingold)
legend(x=-1.1, y=-1.1, c("Deviated reviewers","Experts"), pch=21,
       col="#777777", pt.bg=colrs, pt.cex=2.5, bty="n", ncol=1)
```



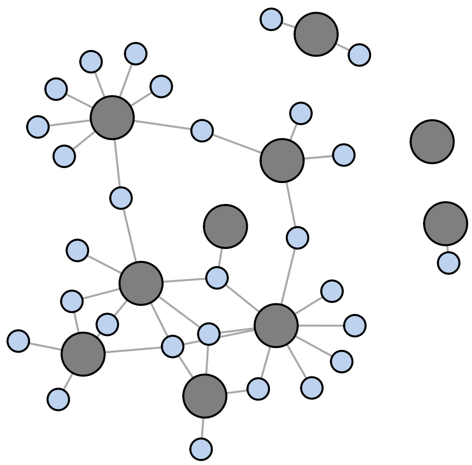
The Kamada Kawai forced-directed algorithm:

```
plot(net,vertex.size=12, edge.arrow.size=0, edge.curved=0,vertex.color=colrs[V(net)$type],
     vertex.frame.color="black",vertex.label=NA,layout=layout.kamada.kawai(net))
legend(x=-1.1, y=-1.1, c("Deviated reviewers","Experts"), pch=21,
       col="#777777", pt.bg=colrs, pt.cex=2.5, bty="n", ncol=1)
```

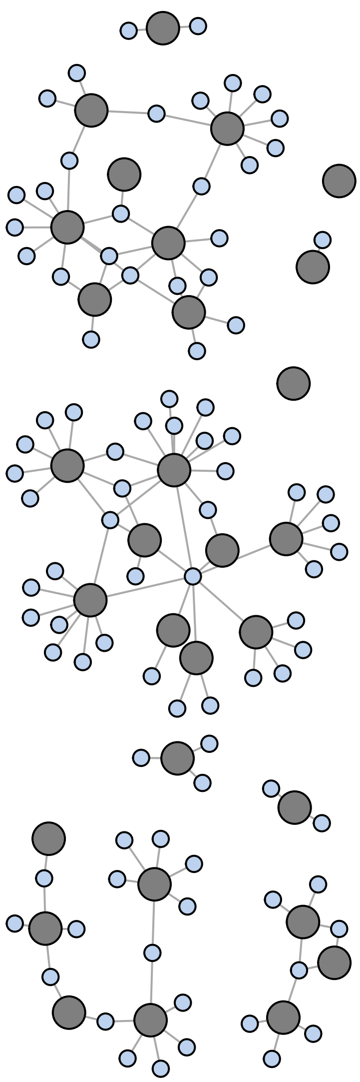


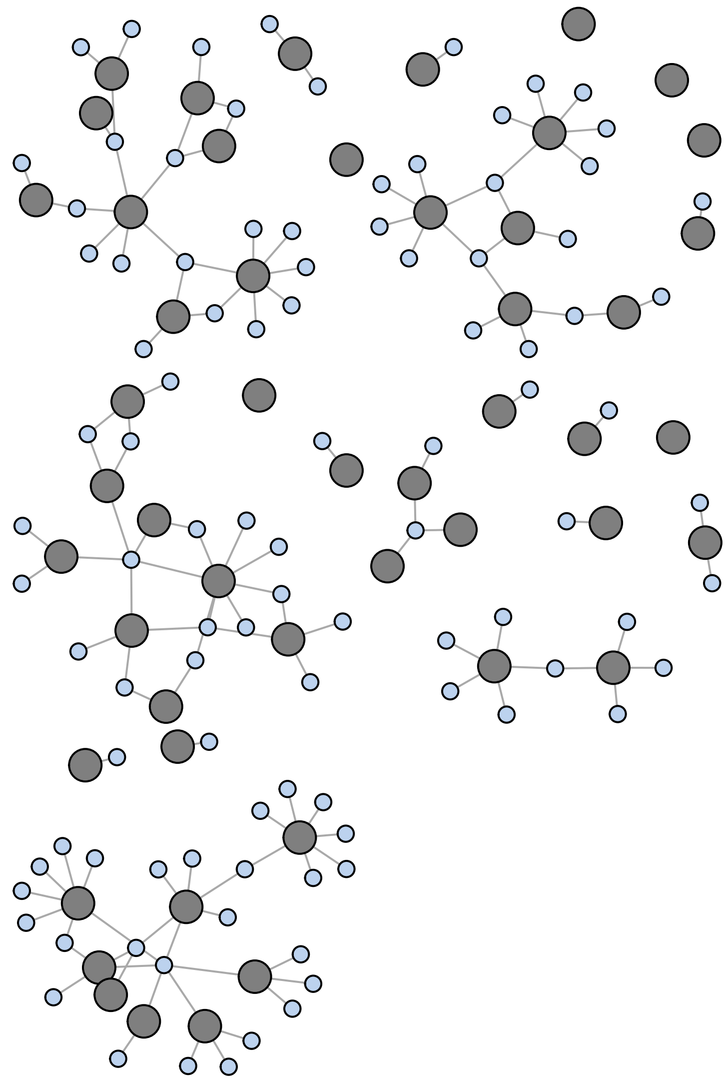
Connect experts with the needed (10 deviated reviewers)

```
colrs <- c("gray50", "lightsteelblue2")
node.new <- node[c(1:10,101:200),]
link.new <- link[which(link$from < 11),]
node.new <- node[c(1:10,unique(link.new$to)),]
net.new <- graph.data.frame(link.new, node.new, directed=T)
net.new <- simplify(net.new, remove.multiple = F, remove.loops = T)
l <- layout.fruchterman.reingold(net.new, repulserad=vcount(net.new)^3,
                                      area=vcount(net.new)^2.4)
plot(net.new, vertex.size=20/V(net.new)$type, edge.arrow.size=0, edge.curved=0,
     vertex.color=colrs[V(net.new)$type], vertex.frame.color="black",
     vertex.label=NA, layout=l)
```



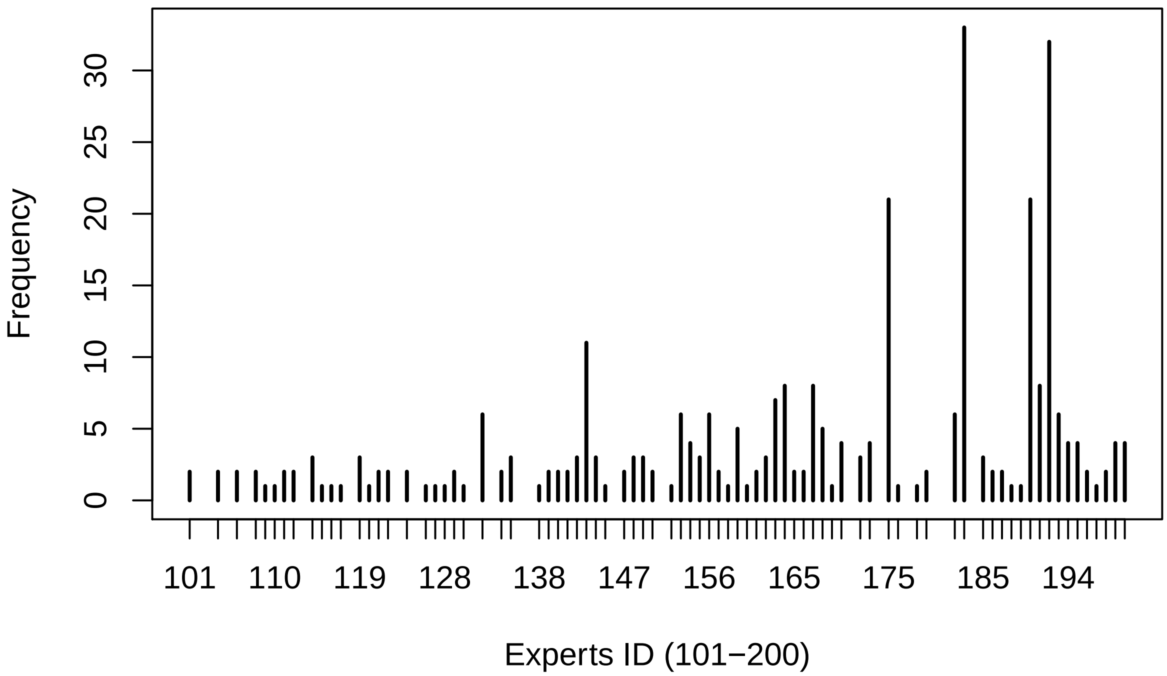
Expert recommendation for all the deviated users:





Exam the involvement of experts in the system:

```r
plot(table(link$to),xlab="Experts ID (101-200)", ylab="Frequency")
```




```r
length(order(table(link$to)))
```

```
## [1] 80
```

```r
#There are 80 experts out of 100 recommended to the deviated reviewers.
#Print 10 most advanced expters
table(link$to)[order(table(link$to))[71:80]]
```

```
##
## 193 163 164 167 191 143 175 190 192 183
##   6   7   8   8   8  11  21  21  32  33
```

Who are they?

```r
adv <- exp[c(93,63,64,67,91,43,75,90,92,83),c(2:6)]
adv
```

```
##    review_num review_ave  help_num help_score       dev
## 93         96   4.416667  7.375000  0.7293581 0.3307292
## 63        223   4.654709 17.699552  0.7539035 0.2780722
## 64         70   4.371429 14.600000  0.7800065 0.2793805
## 67        238   4.689076 35.180672  0.8531945 0.2890479
## 91         64   4.609375 12.078125  0.8715278 0.3287300
## 43         54   4.333333 11.333333  0.9220779 0.2384003
## 75         64   4.718750 10.031250  0.7560153 0.3021759
## 90         59   4.440678 28.474576  0.7509383 0.3286450
## 92        406   4.349754 20.460591  0.7909175 0.3297312
## 83         76   4.197368  7.855263  0.8715479 0.3152513
```

```r
colMeans(adv)
```

```
##  review_num  review_ave    help_num  help_score         dev
## 135.0000000   4.4781138  16.5088363   0.8079487   0.3020163
```

```r
exp_sub <- exp[,c(2:6)]
colMeans(exp_sub)
```

```
##  review_num  review_ave    help_num  help_score         dev
## 125.9300000   4.5914989  15.6489400   0.8000546   0.2402854
```

1. Average number of reviews for movies is considerable high than the experts popylation
   —> No surprise

2. Average review scores for the 10 advanced experts is lower than the experts population
   —> More critical?

3. Deviation of the 10 advanced experts is higher than the experts population
   —> Professional perspective?